View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Directory of Open Access Journals

The current issue and full text archive of this journal is available on Emerald Insight at:
www.emeraldinsight.com/2398-7294.htm

Missing observation approximation for spatio-temporal profile reconstruction in participatory sensor networks

Assad Mehmood
Saudi Ministry of Defense, Riyadh, Saudi Arabia, and
Kashif Zia, Arshad Muhammad and Dinesh Kumar Saini
Department of Computing and IT, Sohar University, Sohar, Oman

Received 30 May 2018
Revised 28 August 2018
Accepted 3 September 2018

Abstract

Purpose – Participatory wireless sensor networks (PWSN) is an emerging paradigm that leverages existing sensing and communication infrastructures for the sensing task. Various environmental phenomenon – *P* monitoring applications dealing with noise pollution, road traffic, requiring spatio-temporal data samples of P (to capture its variations and its profile construction) in the region of interest – can be enabled using PWSN. Because of irregular distribution and uncontrollable mobility of people (with mobile phones), and their willingness to participate, complete spatio-temporal (CST) coverage of *P* may not be ensured. Therefore, unobserved data values must be estimated for CST profile construction of *P* and presented in this paper.

Design/methodology/approach – In this paper, the estimation of these missing data samples both in spatial and temporal dimension is being discussed, and the paper shows that non-parametric technique – Kernel Regression – provides better estimation compared to parametric regression techniques in PWSN context for spatial estimation. Furthermore, the preliminary results for estimation in temporal dimension have been provided. The deterministic and stochastic approaches toward estimation in the context of PWSN have also been discussed.

Findings – For the task of spatial profile reconstruction, it is shown that non-parametric estimation technique (kernel regression) gives a better estimation of the unobserved data points. In case of temporal estimation, few preliminary techniques have been studied and have shown that further investigations are required to find out best estimation technique(s) which may approximate the missing observations (temporally) with considerably less error.

Originality/value – This study addresses the environmental informatics issues related to deterministic and stochastic approaches using PWSN.

Keywords Regression, Approximation, Environmental informatics, Spatial-Temporal profile

Paper type Research paper

1. Introduction

Participatory wireless sensor networks (PWSN) (Bruke *et al.*, 2006; Miluzzo *et al.*, 2010) is an emerging paradigm where instead of deploying dedicated infrastructures, existing

© Assad Mehmood, Kashif Zia, Arshad Muhammad and Dinesh Kumar Saini. Published in *International Journal of Crowd Science*. Published by Emerald Publishing Limited. This article is published under the Creative Commons Attribution (CC BY 4.0) licence. Anyone may reproduce, distribute, translate and create derivative works of this article (for both commercial and non-commercial purposes), subject to full attribution to the original publication and authors. The full terms of this licence may be seen at http://creativecommons.org/licences/by/4.0/legalcode

International Journal of Crowd Science
Vol. 2 No. 2, 2018
pp. 108-122
Emerald Publishing Limited
2398-7294
DOI 10.1108/IJCS-05-2018-0009

sensing (handheld devices) and communication infrastructures (i.e. Cellular, Wifi, WiMax, internet) are used to perform the sensing task. The ubiquity of mobile phone handsets is the key driver of the above mentioned paradigm. In addition to video-image (camera) and acoustic (microphone) sensors, one may attach other sensors (using Bluetooth) to the handset. According to Kansal *et al.* (2007) can enable various applications including noise (and, or) air pollution monitoring, item price sharing, locating urban area parking lots, etc.

First community noise mapping application is considered. Noise pollution is a serious concern in urban areas. Noise maps of some cities (e.g. the UK) available on the Web are generated using computer-based modeling techniques; therefore, they do not precisely represent the actual noise level. Santini *et al.* (2008) have proposed noise pollution monitoring using dedicated sensors, but deploying a large network of dedicated nodes in a city may not be possible. Alternatively, a highly scalable way is to use mobile phones as acoustic sensors when exposed to open air (Kanjo, 2010). The noise-level samples acquired can be relayed to a central data sink to build so-called spatio-temporal noise profile that can be mapped on geographic (Google) map to visualize noise pollution of a given community. Various noise level indicators are quarried from the built profile including $L_{eq,T}$ (equivalent noise level for time period T), $L_{(Ti)}$ (noise level at time interval i), $L\text{per cent}_{,1h}$, $L\text{per cent}_{,day}$, L_{day}, $L_{evening}$. The profile may serve many purposes such as selecting quite place to live in the city, trend analysis of noise pollution, urban planning, help in selecting location (for a new entertainment center, pub, industry, hospital or residential block), or to decide whether to hire acoustic engineers for the survey of the chosen location. The basic design principle of PWSN is the user's contribution (i.e. people carrying mobile devices and willing to participate for sensing task). Furthermore, irregular geographic distribution, unpredictable mobility of people and lack of synchronization of acquired data samples result in loss of coverage. The irregular spatio-temporal sampling and nature of P make the problem significantly different from the missing data problems in other data acquisition systems. Mobile phones are not meant to be used for dedicated sensing, thus negotiating sampling rates (and intervals) with heterogeneous set of users may not work here. Therefore, to construct the complete spatio-temporal (CST) profile of the phenomenon P, estimation of missing data samples is required such that the data remain valuable in spatio-temporal context.

The study aims to investigate the noise data in environmental informatics which is collected while logging from various sensing devices. It has been observed that there is a possibility of some missing data or some incoherent noise intervention at the time of collecting environmental data from different types of logging devices. Attempt has been made to minimize the noise and measure the appropriate missing values. PWSN have been used for measurements and large amount of homogeneous data have been retrieved during the logging process. The study proposes spatio-temporal and geo-statistics mechanisms of data analysis for the real-time cleaning, filtering and mapping. Environmental data informatics is an emerging discipline which helps to clean environmental data. This study addresses the environmental informatics issues related to deterministic and stochastic approaches using PWSN. The focus is on the comparative study of parametric and non-parametric regression techniques to estimate the missing data samples.

The main focus is on the comparative study of parametric and non-parametric regression techniques to estimate the missing data samples. The simulation results show that the non-parametric approach outperforms the parametric techniques thus suits well for estimation task in PWSN. Noise pollution monitoring is used as an illustrative example; however, same estimation techniques are applicable to other applications such as road traffic monitoring, air pollution monitoring and CO_2 monitoring.

The rest of the paper is organized as follows. Section 2 gives an overview of the related work; Section 3 discusses the noise profile construction problem and proposes a solution with a brief introduction to parametric and non-parametric techniques, i.e. deterministic and stochastic approaches along with introduction to the temporal regression techniques. Section 4 describes a prototype implementation of smart phone handset as acoustic sensor. Section 5 presents the simulation model and estimation results. Finally, Section 6 details the findings of the study and provides the future research directions.

2. Related work

The main focus of the current research in WSNs is to use dedicated infrastructures (Akyildiz *et al.*, 2002). An alternative paradigm for WSNs, i.e. PWSN has emerged recently. A few recently proposed projects include Participatory Sensing (Bruke *et al.*, 2006), SenseMart–Sensing Data Market (Chou *et al.*, 2007), Mobile GeoSensing (Kanjo, 2010), MetroTrack (Ahn *et al.*, 2010), SoundSense (Lu *et al.*, 2009) and Bubble-Sensing (Lu *et al.*, 2010). PWSN has also been introduced by some other groups of people including Schweizer *et al.* (2011, 2012), Maisonneuve *et al.* (2010) and Baykasoglu *et al.* (2016) for participatory noise map generation.

Missing (or unobserved) data refers to the difference between data planned to be collected and data collected in actual. This problem has been widely studied in different fields such as statistics, databases, field of medicine (Lakshminarayan *et al.*, 1999; Longford, 2005; Little and Rubin, 2014). Some of the literature about data sample estimation in sensor networks is briefly discussed as under.

There have been number of attempts to deal with missing data in WSN. Jiang and Gruenwald (2007) and Nag *et al.* (2015) propose data estimation technique using association rule mining in the data streams. The problem with these techniques is that the missing behavior in PWSN is mainly because of non-availability of data producing nodes; therefore, association rule mining (Indira and Kanmani, 2015) approaches may not always work which is suitable for data streams with comparatively small-scale missing data. Ensemble techniques have also been used for handling missing data (Mohammed *et al.*, 2006). The ensemble approaches work well when data are corrupt and use weighted majority approach for classification but are not able to handle situations with a large number of missing features and hence work only when limited number of sensors malfunction.

Elnahrawy and Nath (2004) have proposed an online learning of spatio-temporal correlations and utilized them to discover outliers, approximate missing values and detect faulty sensors, however, the assumption behind their work is dense network with redundant data and correlated readings for coverage and connectivity.

Jiang and Gruenwald (2007) and Sutha and Dhanaseelan (2017) proposed the use of association rule mining for estimation in data streams. However, missing data behavior in PWSN (because of non-availability of users), association rule mining approaches (suitable for small-scale missing data) may not always work. The authors proposed spatio-temporal correlations learning to discover outliers, approximate missing values and detect faulty sensors assume a dense network with redundant readings for coverage and connectivity. To build Worldwide Sensor Web, Balazinska *et al.* (2007) have raised various data management issues including missing data and used interpolation to fill them. Guestrin *et al.* (2004) have proposed parametric regression to model sensor data. Deshpande and Madden (2006) have also used the aforementioned approach to estimate missing data values. Although parametric model may work well, however, the approach may not capture unforeseen situations in dynamic environments (Christin *et al.*, 2011). Therefore, fixing a model may result in loss of flexibility to explore, analyze, and decide based on available real-time data.

Therefore, a flexible, data dependent (nonparametric) method to estimate missing samples exploiting their relationship with the sensed data has been proposed (Xu *et al.*, 2016).

3. Sensor coverage and spatio-temporal profile reconstruction

The environmental noise is highly dynamic and fluctuates depending upon nature (frequencies and loudness) of the sound sources (Bies and Hansen, 2009). Typical sound sources include transportation systems (road, rail and air traffic), construction work, factory and audio entertainment systems. To capture most of the noise variations, finely grained sampling is required. Therefore, to acquire noise samples at grid points, entire region of interest can be divided into a regular grid with certain dimensions (say 5 × 5 m). Similarly, the temporal granularity can also be fixed at regular intervals to build CST noise profile of the region.

As in PWSN, granular sampling from the whole region may not be possible; therefore, a mechanism is required to estimate the unobserved data points. The focus is to construct spatial profile for unit time t_i. Once constructed, the profile may be used to estimate unobserved data points at the next time stamp t_j along with the observations collected from the region at time t_j. The preliminary results are provided for the estimation task in time dimension with the point of view that historical data sets may help to reduce the estimation error while approximating an unobserved data point.

The attributes (A) in each data sample S_i, i.e. geographic location $x \in \mathbf{R^2}$ (or $x \in \mathbf{R^3}$ for 3D case); t, time interval; equivalent noise level, $L_{eq,t}$, during t; and time stamp can be represented as:

$$S_j \rightarrow A_j = (a_1 \quad a_2 \quad \ldots \quad a_m). \tag{1}$$

Received data samples at an instance I can be represented as:

$$I = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ a_{11} & a_{12} & \ldots & a_{1m} \\ \vdots & & & \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix}$$

Assume expected number of samples, N_{exp} required for the profile construction of P (based on spatial granularity requirement of P) is known, and N_{recv} be the number of samples received, missing data samples problem can be expressed as: $N_{recv} < N_{exp}$. Thus, completeness of the spatial profile (matrix I) can be determined using:

$$r = N_{recv}/N_{exp} \tag{2}$$

Where $r \in \mathbf{R} \wedge 0 \leq r \geq 1$ assuming $N_{recv} \leq N_{exp}$ (always); hence, the ratio r depicts quantitative measure of missing samples as shown in Figure 1.

The objective of the estimation task is to approximate the missing entries in matrix I using regression techniques as mentioned in Algorithm 1 (Figures 2 and 3).

3.1 Community noise map and estimation techniques

Referring to the noise map construction and noise (sound) propagation in an ideal case follows sound propagation obeys Inverse Square Law Bies and Hansen (2009). The law

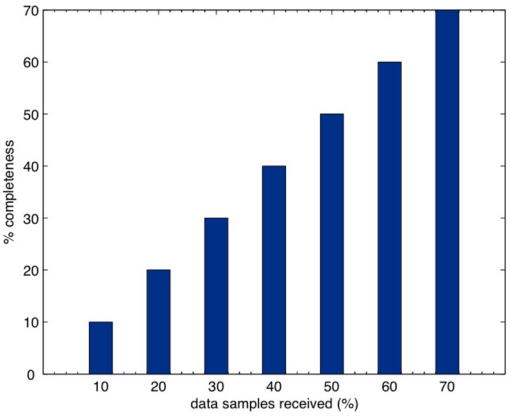

Figure 1.
Data samples received effecting profile completeness

Data: $I_{m,n}$, N_{exp}, N_{recv} , S, such that I contains missing data samples (rows)
Result: $I_{m,n}$ with estimated missing samples
begin
  if $N_{recv} < N_{exp}$ **then**
    $N_{miss} \leftarrow N_{exp}$ - N_{recv}
    for $i \leftarrow 1$ *to* N_{miss} **do**
      foreach *attribute* a *in* S **do** $a \leftarrow$ ESTIMATE(a)
      Add a to S_{est}
      Add S_{est} to I
end
function ESTIMATE(a)
$a_{est} \leftarrow ApplyNWKR(Eq.10)$
$a_{est} \leftarrow ApplyARMA$
return a_{est}

Figure 2.
Algorithm: missing data samples(s) estimation

states that in a "free field", sound will fall at the rate equal to the inverse of square of the distance (from the source). In fact, there are many factors which affect the spherical propagation of sound. As a result, the sound signal attenuates is not allowed to dissipate as Inverse Square Law suggests. The outdoor sound propagation depends on several factors which include refraction (because of gradients of wind and temperature), reflection (at ground, building, forests etc), diffraction (at obstacles e.g. buildings, hills, big signboards/screens etc), scattering (because of atmospheric turbulence), absorption because of humidity. The equation for sound signal propagation is given as:

$$L_p = L_w - 20logr - 11 \quad (dB) \tag{3}$$

where L_p is the sound pressure level, L_w is sound power level and r is the distance from the source in meters. After taking in to account the above-mentioned factors affecting the sound propagation, equation (3) becomes:

$$L_p = L_w - 20logr - 11 - A_{abs} - A_E \quad (dB) \tag{4}$$

where: A_{abs} is atmospheric absorption and A_E is excess attenuation (dB). The total attenuation A_E (dB) is a combination of all effects:

$$A_E = A_{weather} + A_{ground} + A_{turbulence} + A_{barrier} + A_{vegetation} + \ldots \tag{5}$$

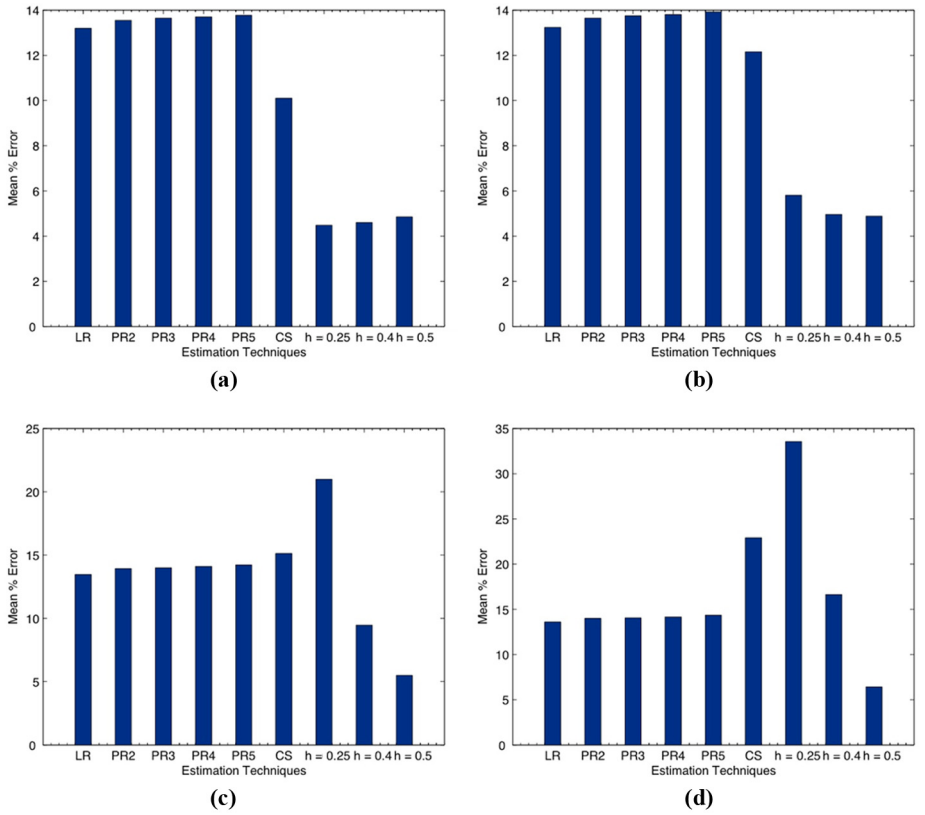

Figure 3.
Regression analysis: parametric vs non-parametric (Kernel) regression

Notes: (a) 10% unobserved data points; (b) 30% unobserved data points; (c) 60% unobserved data points; (d) 70% unobserved data points

These terms are quantified using different set of equations, which can be studied in detail Dunn *et al.* (2015). From the above-mentioned sound propagation model, it is obvious that accurate measurement of all the factors involved in the sound propagation is very difficult, if not impossible, to calculate noise level at certain location r from the source as to exactly measure all the factors involved.

3.2 Estimation in spatial dimension

The inability to acquire maximum (desired) coverage in the data collection (of noise level samples) leads to insufficient and missing data problem as the valuable data required to build the spatio-temporal profile of the whole region may not be available at required resolution. As discussed earlier, to maximize the coverage from the data collected, estimation mechanism is required to estimate the noise level at the locations where data have not been collected. For the significance of the estimation, the approximation procedure should take into account the data collected from other locations. Given that the noise level data are missed at certain location, in fact, there are two ways to estimate the noise level: deterministic and stochastic.

For the deterministic solution, all the parameters should be known which have been discussed above along with the sound pressure level of the sound source(s). The other factors critical for deterministic solution are types (point, cylindrical or plane) and nature (uni or omni directional) of the sound source(s) responsible for propagating noise in the region. By conducting experimental studies, it can be concluded that from given set of noise level data from users, it is complicated procedure to determine the noise level of the sound source(s), and it is hard to determine the number of sound sources involved. Even if it is known, there are many other factors like type and nature of sound source(s), factors involved in sound propagation, etc. that need to be determined to calculate noise level at a given location, which is hard to get in real-world scenarios. Therefore, to determine noise level at a given location using deterministic approach is hard. This leads the study to more realistic approach to estimate the missing noise level data at a given location, also known as stochastic (non-parametric) approach. According to this approach, the noise level at a given location can be estimated based on the data gathered from the neighboring locations. Regression analysis is applied using both *parametric* and *non-parametric* techniques for estimation task. The techniques are briefly discussed below.

3.2.1 Parametric regression. In parametric regression, the aim is to find the best-fit equation for the data, e.g. in case of Linear regression, the relationship between dependent (response, unknown) and independent (explanatory, received data) variables are assumed to be linearly defined:

$$y_i = ax_i + b + \epsilon_i, \qquad i = 1, \ldots, n \tag{6}$$

where y_i is the response, x_i is the explanatory variable, ϵ_i is random error, a is slope and b is the y-intercept of the regression line George and Collins (2003). Equation (6) describes a line in x-y plane with n pairs (x_i, y_i) of data samples on that plane.

The response variable may depend on a nonlinear function of explanatory variables (e.g. realistic carbon emission predicted future global warming is not expected to be a simple linear function of time). We used Polynomial regression to estimate missing data values (George and Collins, 2003).

$$y = f(x) + \epsilon \tag{7}$$

where $f(x)$ is the polynomial (aka basis) function and can be of order 1, 2, 3, ... n. We have also used Cubic spline interpolation; details can be seen in Deboor's (1978) study.

3.2.2 Non-parametric regression. In contrast, non-parametric methods can be used to build up an overall model of the data based on simple and intuitive local models, which can be used for the scenarios with unknown (or difficult to fit) theoretical models (Hardle, 1990). The aforementioned approach suits to the PWSN estimation problem because:

- It allows to tailor estimation problem to the local characteristics of the observed data.
- It is more flexible (and robust) in setting up model parameters dynamically and does not require the data to follow equally spaced sampling.

Kernel regression is used for non-parametric estimation of unobserved data samples, as the method is simple to analyze and is mathematically well understood (Hardle, 1990). The key idea is to use weighted average of the observed data values, where at a given location $x \in \mathbf{R}^2$ (or $x \in \mathbf{R}^3$), noise level at x is conditionally independent of far away locations, defined as:

$$\hat{m}(x) = 1/n \sum_{i=1}^{n} W_i(x) v_i \tag{8}$$

where $\hat{m}(x)$ is the estimated noise level at x, $\{W_i(x)\}_{i=1}^{n}$ denotes sequence of weights that may depend on all the observations at $\{x_i\}_{i=1}^{n}$ and v_i is the noise level observed at x_i. For W, the weighting procedure as proposed by Nadaraya–Watson (Nadaraya, 1964), was applied and thus Nadaraya–Watson Kernel Regression (NWKR) is given as:

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x - x_i)v_i}{\sum_{i=1}^{n} K_h(x - x_i)} \tag{9}$$

where $K_h = K\left(\frac{x-x_i}{h}\right)$ is the kernel function and h is the neighborhood bandwidth of x used to assign weights to the neighboring data samples. Hardle (1990) has shown that h effects the estimation accuracy, therefore, needs to be carefully selected. For Kernel function K, we defined it to be Gaussian, as the choice of K has less impact on estimation accuracy (Hardle, 1990). Gaussian Kernel is defined as:

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\left(-u^2/2\right)}, \text{where } \mathrm{u} = (x - x_i)/h \tag{10}$$

3.3 Estimation in temporal dimension

In this section, the temporal profile reconstruction has been discussed. The data samples acquired can be represented as time-series since the data samples are gathered at successive time stamps for a time period. It is worth mentioning that the static analysis of the acquired data samples for a given time period, i.e. the techniques are applied to the data samples acquired once sampling for the entire time period has finished. For the estimation of missing data values, we have used the popular time-series prediction models such as moving average and autoregressive (AR) models. The purpose of the study is to determine the accuracy of the time-series prediction models and their applicability in the context of participatory sensing, where data samples from a given region may not be synchronized with each other spatially and temporally.

3.3.1 Moving average. Moving average is most commonly used smoothing techniques, which is often used to provide time-series prediction. The techniques work by reducing the randomness of the (time-series) data samples, and may lead to better prediction. Moving average model of order q can be given as under:

$$X_t = \mu + \epsilon_t + \sum_{i=1}^{q} \Theta_i \epsilon_{t-i} \tag{11}$$

where $\Theta_1, \ldots, \Theta_q$ are the parameters of the model, μ is the expectation of X_t and $\epsilon_t, \epsilon_{t-1}, \ldots,$ are the noise error terms.

As in general, time series analysis, simple moving average models performed better than higher window size models. The question arises that, do other advanced time-series prediction models perform any better than moving average models for our estimation problem. To investigate this issue, we evaluate the performance of advanced time-series prediction models. In particular, we investigate the performance of the AR and AR moving average (ARMA) models.

3.3.2 Autoregressive model. The AR model considers a value at time t based on the linear combination of prior values (i.e. forward prediction) and upon the combination of subsequent

values (i.e. backward prediction). In other words, the AR model uses the acquired data values at a particular grid location to predict the values at next time interval. Equation (12) refers to mathematical representation of the AR model of order p.

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \epsilon_t \tag{12}$$

where $\phi_1, \ldots, \phi_p$ are parameters, c is constant and the random variable ϵ_t is the noise.

3.3.3 Autoregressive moving average model. Using both moving average and AR models, ARMA model may also be used for the estimation task. In other words, the model contains both $AR(p)$ and $MA(q)$ models. The mathematical notation of the ARMA model is given in equation (13).

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \phi_i X_{t-i} + \sum_{i=1}^{q} \Theta_i \epsilon_{t-i} \tag{13}$$

Where, c is constant, and the variable ϵ_t is the noise.

4. Acoustic sensor implementation

To sense the noise level at a given location a program in Microsoft Visual C# was developed for hand held smart phones – PDA's to find noise level (dB) at a location for the sound signals captured using built-in microphones. The implementation involves calculation of noise exposure level (L_E, a metric to measure noise level, with reference time interval set to 1 second) as given below:

$$L_E = 10log_{10}\left[\frac{1}{t_0}\int_{t_1}^{t_2}\frac{p^2(t)}{p_0^2}dt\right] \tag{14}$$

As discussed above, the implementation involves calculation of the noise exposure level L_E as per equation (14), where reference time interval can be set up to the desired number. Currently, the reference time interval has been set to 1 s. The implementation involves various steps, which are briefly stated as under.

By using a set of library functions, which give access up to microphone level for PDA's (Mitchell, 2007), the sound samples were acquired to determine noise dB level. As microphone converts the sound signals into electrical signals, the energy level of the captured sound samples was calculated by applying fast Fourier transform, followed by a weighting filter to assign weights to the transformed signals and then Parseval's relation (Mitra and Kuo, 2006), which is used to estimate the signal energy as shown in equation (15).

$$E = \sum_{n=0}^{N-1} (x[n])^2 \tag{15}$$

The final dB(A) is calculated using the relationship:

$$\text{Signal Level in dB(A)} = 10log_{10}\left(\frac{E}{E_{ref}}\right) \tag{16}$$

where E_{ref} is the (constant) reference signal level, equation (16) becomes:

$$\text{Signal Level in dB(A)} = 10log_{10}E + C \tag{17}$$

where C is calibration constant and can be determined by acquiring samples and cross reference with the sound level meter readings.

5. Simulation model and results

MATLAB was used to simulate noise profile construction scenario from the sparse sampling to study the accuracy of various regression techniques. Vehicular traffic noise propagation (Makarewicz, 1998) model was considered to disperse the noise from the center of the main road carrying metropolitan traffic (over 1500 vehicles per hour in the peak period). Vehicles on the road follow Poisson arrival rate (Songchitruksa and Hard, 2008). The sensing nodes (people) are positioned at grid of 1 m size on the sidewalk along the road (of dimensions 100 × 4 m). The ground truth (noise level) was gathered at all the grid points. Moreover, different sparse data sets were generated from the ground truth by randomly omitting the samples from grid points at certain (increasing) percentage e.g. 10, 20 and 30 per cent.

It is worth noting that realistic model has been used for environmental noise propagation, as experimental settings on even small-scale require quite a few resources and the results might get affected by environmental and or human influences. However, in the next step the empirical study is supposed to be done to compare our findings with simulation results.

5.1 Profile reconstruction in spatial dimension

This section discusses the simulation results for the estimation in the spatial dimension. The regression techniques were implemented, and applied on the generated sparse data sets to estimate the omitted (unobserved) traffic noise data samples.

To cross-validate (with the ground truth) and measure the accuracy of the estimation results, the standard metric mean absolute percentage error (MAPE) (Hardle, 1990) was used for the 100 runs (of each technique) with randomly omitted samples. Figure 6 shows the comparison results of various regression techniques i.e. Linear Regression (*LR*), polynomial regression of degree 2 to 5 (*PR*2, . . ., *PR*5), Cubic Spline (*CS*) and Gaussian Kernel regression (with 0.25, 0.4 and 0.5 meters of h). The data sets used for the comparison contain 10, 30, 60, and 70 per cent unobserved points respectively. The results shows that the kernel regression outperforms the parametric regression techniques, because of the fact that modeling P with real-time unpredictable behavior with fixed parametric models is hard, therefore, non-parametric models should be preferred.

In further simulations, it was observed that MAPE is inversely proportional to h [shown in Figure 4(a)] because, for the higher value of h, more data points are involved for the estimation and the weights of the observed data points are smoothed decreasing the overall MAPE. It can be seen that increase in MAPE up to a certain percentage of unobserved data points is not very significant because grid points are closely located and noise variation is less because of single type of sound source being used. Figure 4(b) presents 90 per cent confidence interval of MAPE that increases with the increase in the percentage of unobserved data points.

5.2 Profile reconstruction in time dimension

The experimental results for the temporal profile reconstruction using the techniques discussed above 3.3 are discussed here. Upon analyzing the time series data for the grid points for the current simulation scenario, results for increasing percentage of unobserved

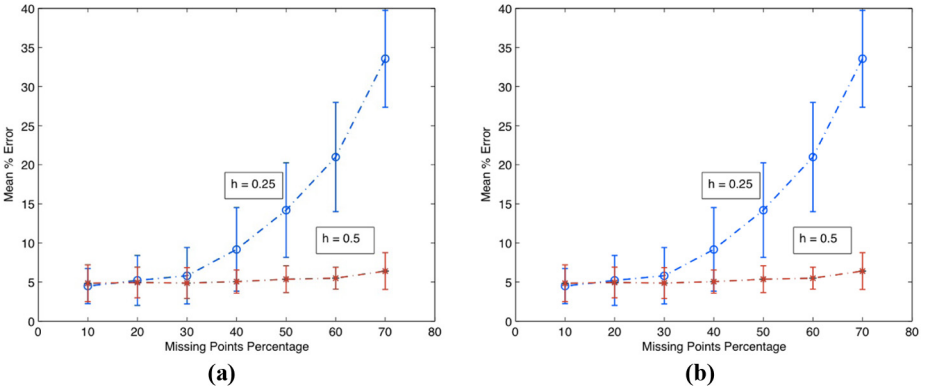

Figure 4.
Guassian Kernel regression analysis

Notes: (a) Guassian Kernel Regression (NWKR); (b) NWKR, MAPE Confidence Interval

(missing) data values suggest that moving average has failed to model the series, as indicated by the accuracy metric, i.e. mean percentage error (MPE) aka *mean prediction error*. The results can be generalized as shown in the Figure 5 for this particular simulation scenario, as the MPE for moving average converges for the data samples, where each data sample location is the grid point location in the simulations.

Figure 6(a) plots the mean prediction error as a function of percentage of missing data samples, for different moving average window sizes. The increase of missing data samples results in a linear increase of the MPE. For low percentage, missing data the higher window sizes of moving average have the ability to considerably reduce the MPE, whereas, at the higher percentage of missing data, the reduction of MPE is less. To better describe the effect of increasing the window size of the moving average at higher missing data samples, Figure 6(b) presents the MPE as a function of moving average window sizes. It is evident that for lower percentage of missing data samples, the simple moving average models

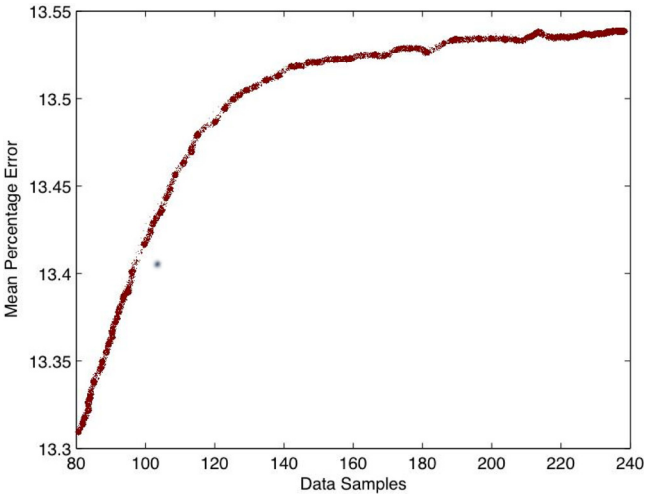

Figure 5.
MPE for increasing data sample size

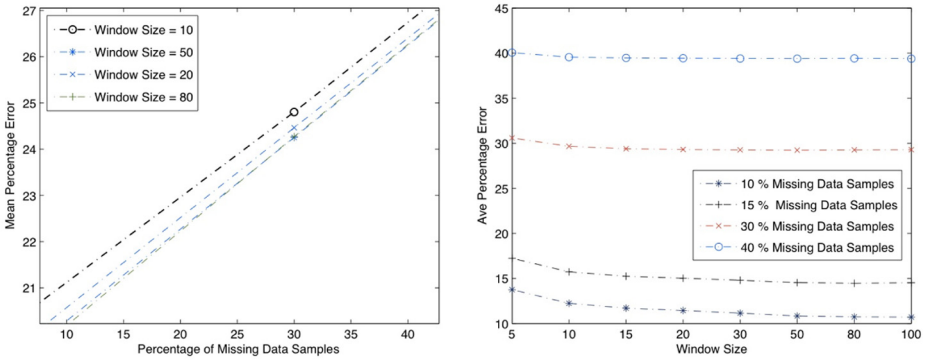

Figure 6.
Moving average

Notes: (a) MPE as function of unobserved data points; (b) MPE as function of window size

(e.g. moving average with window size of 5 and 10) reduce the MPE of prediction. However, higher window size models have no significance in reducing the average percentage error.

Figure 7(a) plots the mean prediction error as a function of missing data samples, for different orders of AR model. As the case with moving average models, increasing the percentage of missing data samples results in a linear increase of the mean prediction error. However, higher-order AR models increase the MPE. The results are found surprising, as it suggests that the higher-order AR models are not fitting the data under investigation, which may point out to the fact that the correlation between the different points of the time-series does not carry any significance. To conclude this, further investigations are required and will be done as part of the future work. Figure 7(b) is plotted to further investigate the effect of increasing the order of AR model on mean prediction error, the figure confirms that the higher-order AR models increases the MPE.

Although the synthetic data generated by the vehicular propagation model is expected to be noise-free, yet the performance of ARMA models is being investigated. ARMA models usually applied in the cases where data are affected by noise, as it enhances the prediction of the AR model by incorporating the noise in the estimation process. Figure 8(a) and (b)

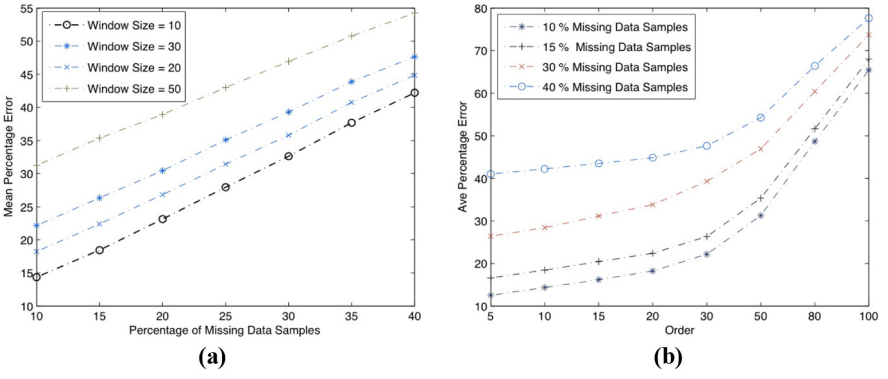

Figure 7.
AR model

Notes: (a) MPE as function of order; (b) MPE as function of unobserved data points

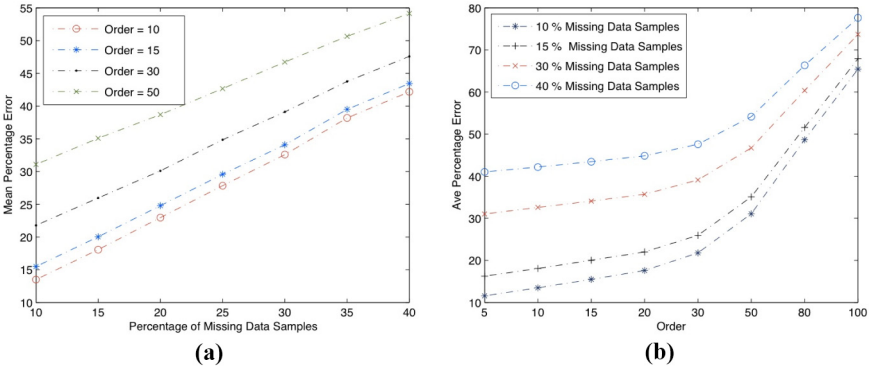

Figure 8.
ARMA model

Notes: (a) Mean percentage error as function of order; (b) mean percentage error as function of unobserved data points

clearly confirms that the results of ARMA are very similar to the results gathered by the AR model, which emphasizes that the data is noise-free.

6. Conclusion and future work

The estimation of unobserved data samples for PWSN applications, especially *environmental noise pollution monitoring*, in particular, has been discussed. Acoustic sensor implementation and practical utilization of microphones as acoustic sensor for the smart devices has also been discussed. Various estimation techniques for both spatial and temporal profiles have been used to approximate missing observations, and the results for the techniques have been compared. The basic model for expected, received and missing samples provided. For the task of spatial profile reconstruction, it is shown that non-parametric estimation technique (kernel regression) gives a better estimation of the unobserved data points. In case of temporal estimation, few preliminary techniques have been studied and have shown that further investigations are required to find out best estimation technique(s), which may approximate the missing observations (temporally) with considerably less error. This task will be done as part of any future study. Furthermore, in the future study the following will be investigated:

- Integration of spatial and temporal profile estimates to produce CST (urban noise) profile in a region, as both spatial and temporal observations seem to correlate in other words, may assist each other for better approximations.
- The approximation algorithm developed here will be implemented and further improved on more realistic scenarios, (as in reality it is hard to fix the location of mobile users at the grid points) to find out how the proposed scheme works in a realistic environment.

References

Ahn, G.S., Musolesi, M., Lu, H., Olfati-Saber, R. and Campbell, A. (2010), "Metrotrack: predictive tracking of mobile events using mobile phones", in *International Conference on Distributed Computing in Sensor Systems*, pp. 230-243.

Akyildiz, I.F., Su, W., Sankarasubramaniam, Y. and Cayirci, E. (2002), "Wireless sensor networks: a survey", *Computer Networks*, Vol. 38 No. 4, pp. 393-422.

Balazinska, M., Deshpande, A., Franklin, M.J., Gibbons, P.B., Gray, J., Hansen, M., Liebhold, M., Nath, S., Szalay, A. and Tao, V. (2007), "Data management in the worldwide sensor web", *IEEE Pervasive Computing*, Vol. 6 No. 2, pp. 30-40.

Baykasoglu, A., Kaplanoglu, V. and Sahin, C. (2016), "Route prioritisation in a multi-agent transportation environment via multi-attribute decision making", *International Journal of Data Analysis Techniques and Strategies*, Vol. 8 No. 1, pp. 47-64.

Bies, D.A. and Hansen, C.H. (2009), *Engineering Noise Control: theory and Practice*, CRC press, Boca Raton, FL.

Bruke, J. Estrin, D. Hansen, M. Parker, A. Ramanthan, N. Reddy, S. and Srivastava, M.B. (2006), Participatory sensing, in 'ACM WSW' 06'.

Chou, C.T., Bulusu, N. and Kanhere, S. (2007), "Sensing data market", Proceedings of Poster Papers p. 13.

Christin, D., Reinhardt, A., Kanhere, S.S. and Hollick, M. (2011), "A survey on privacy in mobile participatory sensing applications", *Journal of Systems and Software*, Vol. 84 No. 11, pp. 1928-1946.

Deboor, C. (1978), A Practical Guide to Splines.

Deshpande, A. and Madden, S. (2006), "Mauvedb: supporting model-based user views in database systems", *in* '*Proceedings of the 2006 ACM SIGMOD international conference on Management of data*', *ACM*, pp. 73-84.

Dunn, F., Hartmann, W., Campbell, D. and Fletcher, N.H. (2015), *Springer Handbook of Acoustics*, Springer.

Elnahrawy, E. and Nath, B. (2004), "Context-aware sensors", *in 'European Workshop on Wireless Sensor Networks*', Springer, pp. 77-93.

George, W. and Collins, I. (2003), Fundamental Numerical Methods and Data Analysis.

Guestrin, C., Bodik, P., Thibaux, R., Paskin, M. and Madden, S. (2004), "Distributed regression: an efficient framework for modeling sensor network data", *in* '*Information Processing in Sensor Networks, 2004. IPSN 2004. Third International Symposium on*', *IEEE*, pp. 1-10.

Hardle, W. (1990), *Applied Nonparametric Regression*, Cambrige University Press, Cambdrige.

Indira, K. and Kanmani, S. (2015), "Mining association rules using hybrid genetic algorithm and particle swarm optimisation algorithm", *International Journal of Data Analysis Techniques and Strategies*, Vol. 7 No. 1, pp. 59-76.

Jiang, N. and Gruenwald, L. (2007), "'Estimating missing data in data Streams'", *Advances in Databases: Concepts, Systems and Applications*, Springer, Berlin, Heidelberg. pp. 981-987.

Kanjo, E. (2010), "Noisespy: a real-time mobile phone platform for urban noise monitoring and mapping", *Mobile Networks and Applications*, Vol. 15 No. 4, pp. 562-574.

Kansal, A., Nath, S., Liu, J. and Zhao, F. (2007), "SenseWeb: an infrastructure for shared sensing", *IEEE Multimedia*, Vol. 14 No. 4, pp. 8-13.

Lakshminarayan, K., Harp, S.A. and Samad, T. (1999), "Imputation of missing data in industrial databases", *Applied Intelligence*, Vol. 11 No. 3, pp. 259-275.

Little, R.J. and Rubin, D.B. (2014), *Statistical Analysis with Missing Data*, Vol. 333, John Wiley & Sons, Hoboken, NJ.

Longford, N.T. (2005), *Missing Data and Small-Area Estimation*, Springer, New York, NY.

Lu, H., Lane, N.D., Eisenman, S.B. and Campbell, A.T. (2010), "Bubble-sensing: Binding sensing tasks to the physical world", *Journal of Pervasive and Mobile Computing*, Vol. 6 No. 1, pp. 58-71.

Lu, H., Pan, W., Lane, N.D., Choudhury, T. and Campbell, A.T. (2009), "Soundsense: scalable sound sensing for people-centric applications on mobile phones", *in* '*Proceedings of the 7th international conference on Mobile systems, applications, and services*', *ACM*, pp. 165-178.

Maisonneuve, N., Stevens, M. and Ochab, B. (2010), "Participatory noise pollution monitoring using mobile phones", *Information Polity*, Vol. 15 Nos 1/2, pp. 51-71. available at: http://dl.acm.org/citation.cfm?id=1858974.1858981

Makarewicz, R. (1998), "A simple model of outdoor noise propagation", *Applied Acoustics*, Vol. 54 No. 2, pp. 131-140.

Miluzzo, E., Cornelius, C.T., Ramaswamy, A., Choudhury, T., Liu, Z. and Campbell, A.T. (2010), "Darwin phones: the evolution of sensing and inference on mobile phones", *in* '*Proceedings of the 8th international conference on Mobile systems, applications, and services*', *ACM*, pp. 5-20.

Mitchell, C. (2007), 'Adjust your ring volume for ambient noise', MSDN Magazine.

Mitra, S.K. and Kuo, Y. (2006), *Digital Signal Processing: A Computer-Based Approach*, Vol. 2, McGraw-Hill Higher Education, PA Plaza New York, NY City.

Mohammed, H.S., Stepenosky, N. and Polikar, R. (2006), "An ensemble technique to handle missing data from sensors, in 'Sensors Applications Symposium", 2006. *Proceedings of the 2006 IEEE*', *IEEE*, pp. 101-105.

Nadaraya, E.A. (1964), "On estimating regression", *Theory of Probability and Its Applications*, Vol. 9 No. 1, pp. 141-142.

Nag, B.N., Han, C. and Yao, D-q. (2015), "Information enhancement in data mining: a study in data reduction", *International Journal of Data Analysis Techniques and Strategies*, Vol. 7 No. 1, pp. 3-20.

Santini, S., Ostermaier, B. and Vitaletti, A. (2008), "First experiences using wireless sensor networks for noise pollution monitoring", in '*Proceedings of the workshop on Real-world wireless sensor networks*', *ACM*, pp. 61-65.

Schweizer, I., Bärtl, R., Schulz, A., Probst, F. and Mühlhäuser, M. (2011), "Noisemap – real time participatory noise maps", in '*Proceedings of the Second International Workshop on Sensing Applications on Mobile Phones*', *PhoneSense' 11, ACM, New York, NY*.

Schweizer, I., Meurisch, C., Gedeon, J., Bärtl, R. and Mühlhäuser, M. (2012), "Noisemap: multi-tier incentive mechanisms for participative urban sensing", in '*Proceedings of the Third International Workshop on Sensing Applications on Mobile Phones*', *PhoneSense '12, ACM, New York, NY*, pp. 9:1-9:5, available at: http://doi.acm.org/10.1145/2389148.2389157

Songchitruksa, P. and Hard, E.N. (2008), "Queuing simulation of roadside survey station: Blocked traffic lane", *Transportation Research Part A: Policy and Practice*, Vol. 42 No. 6, pp. 857-873.

Sutha, M.J. and Dhanaseelan, F.R. (2017), "Mining frequent, maximal and closed frequent itemsets over data stream-a review", *International Journal of Data Analysis Techniques and Strategies*, Vol. 9 No. 1, pp. 46-62.

Xu, Z., Zhang, H., Sugumaran, V., Choo, K.K.R., Mei, L. and Zhu, Y. (2016), "Participatory sensing-based semantic and spatial analysis of urban emergency events using mobile social media", *EURASIP Journal on Wireless Communications and Networking*, Vol. 2016 No. 1, pp. 1-9.

Corresponding author
Arshad Muhammad can be contacted at: amuhammad@soharuni.edu.om

For instructions on how to order reprints of this article, please visit our website:
www.emeraldgrouppublishing.com/licensing/reprints.htm
Or contact us for further details: **permissions@emeraldinsight.com**